STAT 201: Elementary Statistics Session 13 & 14 Spring 2015
Exam 2
Instructor: Haigang Liu

1. *(20 points)* `cars` is a well known dataset in R package. The data were recorded in the 1920s. It contains 50 observations on 2 variables (the speed of cars and the distance to stop). We study the relationship, by linear regression, the relationship between Y, the distance to stop, and X, the speed of cars.

1) Read the output from StatCrunch in Figure 1 and and comment on strength, direction of the correlation between X and Y. Is it justifiable to apply a linear model in this scenario? Explain.

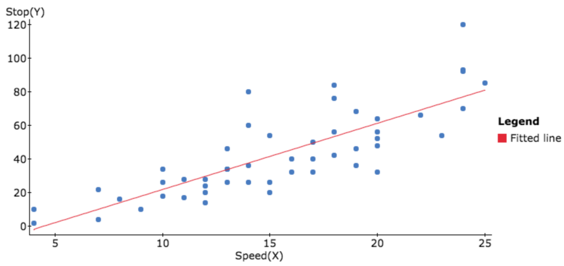

Figure 1: Scatter plot

2) Write down the linear regression model based on the output given by Figure 2.

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	-17.579095	6.7584402	≠ 0	48	-2.601058	0.0123
Slope	3.9324088	0.41551278	≠ 0	48	9.46399	<0.0001

Figure 2: Regression effects

3) Interpret the slope of the regression equation you obtain in 2) *in the context of* this problem.
4) Apply this equation to predict the stop distance for X = 20.
5) Suppose the true value of Y when $X = 20$ is 50. Find the residual of your prediction.

2. *(10 points)* Three PhD students in statistics department with no particular background in driver's education decide to take the permit exam in South Carolina. Each exam is graded as a Pass or Fail. Answer the following questions.

1) Denote the results of their tests as evet X. Find the sample space of X.
2) We assume that all possible outcomes are equally likely. Calculate the probability of all three passing the exam.

3. *(20 points)* The following table indicating a sample of 151 weather events was used for the verification of gale forecasts. Note: a gale is a severe weather event similar to a hurricane. Additionally,

- Let F = gale was forecast
- Let F^c = gale was NOT forecast
- Let G = gale observed
- Let G^c = gale NOT observed

Read the given table in Figure 3, and answer the following quesions.

		Gale observed		Totals
		G	G^c	
Gale Forecast	F	15	2	
	F^c	11	123	
Totals				151

Figure 3: The contigency table of Gale and prediction

1) For this sample, what is the probability that a gale was forecast, i.e., $P(F)$?
2) For this sample, what is the probability that a gale was observed, i.e., $P(G)$?
3) For this sample, what is the probability that a gale was observed AND that it was forecast, i.e., $P(G \text{ and } F)$?
4) For this sample, given that a gale was forecast, what is the probability that a gale actually was observed, i.e., $P(G|F)$?
5) For this sample, are F and G independent events? Justify your answer using probabilities.

4. *(10 points)* How to tell the difference between fabricated data and data from real world? This might be a question posed after the accounting scandals with Enron, a large energy company. In fact, there is a way named Benford's Law to examine accounting books to determine whether they had been "doctored" or used fabricated data. The Benford's law states that, under a variety of circumstances, numbers of population in small towns, figures in newspapers and magazines, tax returns and other business records are more likely to begin with 1. More precisely, the distribution of beginning digits of these numbers of is given in Table 1. Read the table and answer the following questions.
1) Verify that this is a valid probability distribution.
2) Find the expectation of X, i.e, EX.
3) When you randomly select a digit with a random number table, the probability of obtaining 1-9 as starting digit are equally likely. That explains why people are more likely to pick 5 or 6 as starting digits when they fabricate numbers by themselves. Find and compare the probability of 5 or 6 as starting digits by i) random selection and ii) Benford's Law.

5. *(20 points)* A knowledge test is required to acquire driver's permit. In Washington DC, the first part of test has 6 questions for road signs. In each question there are 3 possible answers. Suppose each question is answered by random guess.

x	P(x)
1	0.30
2	0.18
3	0.12
4	0.10
5	0.08
6	0.07
7	0.06
8	0.05
9	0.04

Table 1: The distribution of starting digits given by Benford's Law.

1) The number of right answers is a binomial distribution. Justify this claim by checking the assumptions of binomial distribution.
2) Exam takers will be marked as Fail unless he or she makes all 6 questions correct. Find the probability of passing that test by random guess.
3) Give the mean (expectation) and standard deviation of number of right answers based on random guess.
4) Which of the following graph better describes the shape of the distribution given in 1)? The left one or the right one?

6. *(20 points)* A company makes electronic components for TV's. The probability that a component will fail inspection is 0.1. 240 components are inspected in one day. Let X be the number of components that fail inspection.

1) Find the mean and standard deviation of X.
2) Can we apply emipirical rule in this case? Justify your answer.
3) Find the interval that contain 99.7 % of all observations by applying the empirical rule. In our daily evaluation of the 240 components, we found that 5 of them failed our inspection. Are you surprised by this?

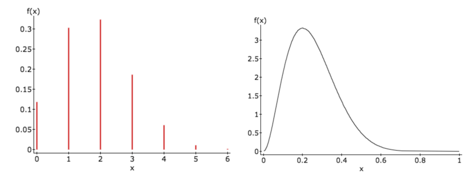

Figure 4: Choose the binomial distribution